Визуализация больших данных с применением методов корреляции событий информационной безопасности в инфраструктуре интернета вещей

С.Ю. Исхаков[1], А.Ю. Исхаков[1], Р.В. Мещеряков[1]
iskhakov.sy@gmail.com | iskhakovandrey@gmail.com | mrv@ieee.org
[1]Томский государственный университет систем управления и радиоэлектроники, Томск, Россия

Статья посвящена вопросам визуализации больших объемов данных, обрабатываемых в процессе обеспечения безопасности инфраструктуры интернета вещей. Рассмотрены методы анализа и обработки больших потоков данных. Проведен обзор и сравнение методов нормализации и корреляции событий в системах управления информационной безопасностью. Предложена методика визуализации как средство совершенствования методов управления инцидентами и представлены результаты ее апробации. Рассмотрены варианты использования современных инструментов для поиска и визуализации больших данных.

Ключевые слова: *интернет вещей, визуализация больших данных, нормализация, корреляция, инцидент.*

Big data visualization with application of correlation methods of information security events in Internet of things infrastructure

S.Y. Iskhakov[1], A.Y. Iskhakov[1], R.V. Mescheryakov[1]
iskhakov.sy@gmail.com | iskhakovandrey@gmail.com | mrv@ieee.org
[1]Tomsk state university of control system and radioelectronics

Article is devoted to questions of visualization of big data processed in the Internet of things infrastructure. Methods of the analysis and processing of big data flows are considered. The review and comparison of normalization and correlation methods in security information management systems is carried out. The visualization technique as means of incident management methods improvement is offered and approbation results are presented. Modern tools for search and visualization of big data are considered.

Keywords: *Internet of things, visualization of big data, normalization, correlation, incident*.

1. Введение

Широкое распространение беспроводных технологий передачи данных, развитие сервисов облачных вычислений и увеличение адресного пространства в сети Интернет посредством внедрения IPv6 [1] стали основными предпосылками возникновения новой парадигмы – Интернета вещей (Internet of Things, IoT) [1], что можно определить как динамическую глобальную сетевую инфраструктуру на основе стандартных и совместимых протоколов связи, где физические и виртуальные «вещи» имеют идентификаторы, используют интеллектуальные интерфейсы и интегрируются в информационную сеть. Обобщение результатов научных исследований, проведенных за последние несколько лет авторским коллективом научной школы, позволяет говорить о глубокой интеграции современных робототехнических систем и IoT-инфраструктуры.

В отличие от компьютеров и смартфонов, значительный процент IoT-устройств не способен применять средства защиты от вредоносного программного обеспечения из-за отсутствия инфраструктуры для запуска подобных приложений. Однако, подавляющее большинство подобных устройств способны генерировать события для ведения журналов. Эта особенность позволяет реализовывать решения по их защите путем анализа лог-файлов. Поскольку в состав IoT-инфраструктуры входит огромное количество устройств с высоким уровнем неоднородности, то генерируемые журналы событий образуют большие объемы данных различного формата.

Настоящая статья нацелена на обобщение методических и практических наработок авторов в части организации сбора и обработки подобных данных для выявления инцидентов информационной безопасности. В качестве средства совершенствования методов управления инцидентами предложена методика визуализации.

2. Сбор и нормализация данных

Общих требований к структуре журналов событий на рынке IoT в настоящее время нет, поэтому разработчики аппаратных и программных продуктов создают системы логирования, исходя из собственных принципов. В одних случаях сообщения записываются в текстовый файл, в других отправляются на syslog-сервер [4,6] или помещаются в базу данных. Форматы хранения могут быть различными, даже если основной механизм транспорта лог-файлов совпадает. Несмотря на это наблюдается переход от средств управления журналами регистрации (Log Management) к SIEM-системам [6], обеспечивающим возможность анализа регистрируемых событий с точки зрения информационной безопасности.

Для событий, полученных от разных источников, могут отличаться формат (TXT, XML, JSON) [2] и стандарт записи. Нормализация – процедура приведения необработанного события к нормализованному виду в соответствии с заранее заданной для источника и типа события формулой нормализации. Два основных этапа нормализации это «разбор событий» (парсинг) и «сопоставление полей» (маппинг). Учитывая большое количество источников данных в инфраструктуре интернета вещей, необходимо осуществлять фильтрацию и укрупнение выборок событий, подлежащих анализу. Для сокращения таких выборок выполняется агрегация – процесс отбора событий, удовлетворяющих условию заранее настроенного правила агрегации, и объединения их в одно агрегированное событие. Ниже представлен пример нормализованного события в MaxPatrol SIEM [2]. В данной системе используется стандарт, разработанный на основе Common Event Expression (CVE) [2,7].

Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18

"action": "allow",
"object": "connection",
"status": "success",
"datafield1": "Access_Control_Policy_DMZ",
"datafield2": "HTTP",
"dst.ip": "192.168.7.16",
"dst.port": 80,
"event_src.category": "IDS/IPS",
"event_src.hostname": "NGIPSV-5555X",
"event_src.title": "asa_firepower",
"event_src.vendor": "cisco",
"importance": "info",
"protocol": "TCP",
"src.ip": "195.38.41.42",
"src.port": 26449,
"time": "2018-04-23T11:20:26Z"

Имея в распоряжении нормализованные и агрегированные данные о состоянии информационной безопасности ИТ-инфраструктуры, можно осуществлять действия, направленные на анализ событий и выявление инцидентов. Основным механизмом в решении данной задачи является применение методов корреляции.

3. Корреляция событий

В рамках данного исследования под инцидентом авторами понимается наличие на отрезке наблюдений зафиксированного факта нехарактерного изменения в сценарии работы объекта (устройства), подтвержденного результатом выполнения условия правила корреляции.

Набор данных, подлежащий обработке правилами корреляции, формируется экспертами на этапе нормализации событий и его описание выходит за рамки данной статьи. В терминах настоящего исследования корреляция событий – это процесс обнаружения инцидентов информационной безопасности путем анализа потока нормализованных данных [2,5]. При обнаружении в потоке событий такой их последовательности, которая указана в условии одного из заранее настроенных правил корреляции, регистрируется корреляционное событие.

В [3,5,7] представлена классификация, которая различает сигнатурные и бессигнатурные методы корреляции. Сигнатурные методы (англ. термин rule based) подразумевают создание человеком неких правил определения инцидентов. Бессигнатурные основаны на обнаружении аномалий по принципу черного ящика, среди которых выделяются подходы, основанные на спецификации и базирующиеся на интеллектуальном анализе данных. Анализ рынка [5-8] SIEM-систем, позволил сформировать список наиболее распространенных к применению на практике методов:

Statistical – бессигнатурный метод корреляции событий, основанный на измерении двух или более переменных и вычислении степени статистической связи между ними.

RBR Rule-based (pattern based) – метод, в котором взаимосвязи между событиями определяются аналитиками в заранее заданных специфических правилах.

CBR Codebook (case based). Корреляция производится по подходящим векторам из предварительно заданной матрицы событий.

MBR (model based reasoning) — метод основан на абстракции объектов и наблюдения за ними в рамках модели.

Graph based. Корреляция определяется на основе поиска зависимостей между системными компонентами в графическом представлении и построении на их основе графа (рис. 1). В случае обнаружения корреляции граф используется для поиска причины инцидента.

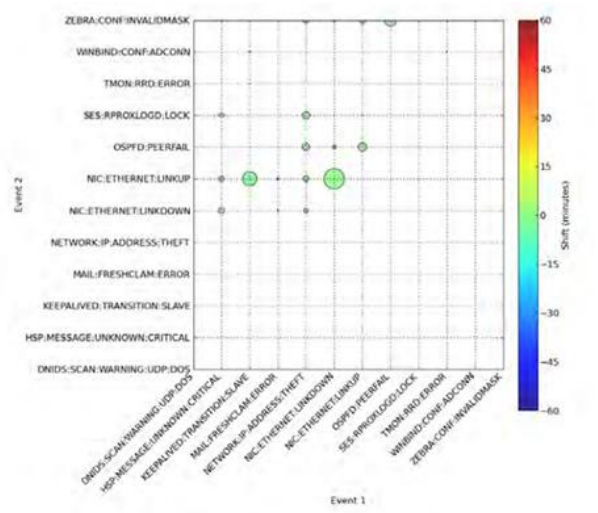

Рис. 1. Пример использования метода Grpah based

В большинстве случаев выявление причины необходимо для адекватной реакции на инцидент или осуществляется в контексте инцидента (например, пользователь изменил параметры запуска служб, необходимо выяснить причину). Наступлению каждого инцидента предшествуют различные события S (рис. 2): сканирование сетевых ресурсов, попытки установить соединение, подозрительные вложения в почтовом трафике. Корреляция позволяет объединить и группировать их для определения момента возникновения инцидента.

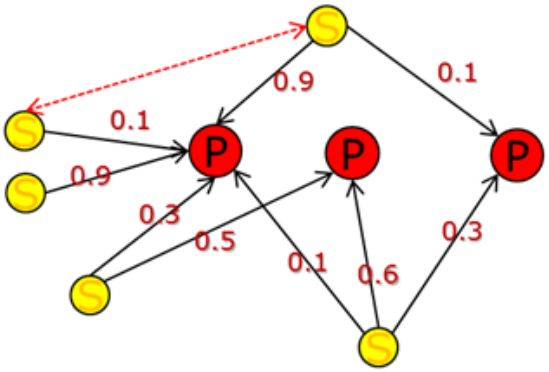

Рис. 2. Использование вероятностной модели

Правило имеет триггер, срабатывающий по условию, счетчик и сценарий реакции. Часть систем включают интуитивно понятные правила в графическом режиме. Счетчики предназначены для подсчета количества совпадений по одному и тому же правилу. Правила корреляции могут включать условия различной сложности.

4. Методика визуализации инцидентов безопасности

Визуализация информации о событиях и инцидентах в инфраструктуре интернета вещей предназначена для интерпретации и принятия решений по оперативной корректировке деятельности при реагировании на инциденты. В [6,7] представлены подходы к формализации данной задачи, в [5] предложена методика для визуализации данных о состоянии топологии сетей различного масштаба. Недостатком данного методического обеспечения является его направленность на визуальное представление топологии сети и состояния сетевых объектов (хостов). Ниже представлена методика визуализации данных, обрабатываемых SIEM-системами с учетом методов нормализации, агрегации и корреляции событий.

The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18

Шаг 1. Определение и выбор ИТ-активов, являющихся источниками данных в рамках конкретной решаемой задачи. Для каждого ИТ-актива, используемого в конкретной задаче необходимо определить механизмы транспорта лог-сообщений (интерфейсы и протоколы взаимодействия) и организовать получение событий SIEM-системой.

Шаг 2. Провести анализ получаемых событий с целью выделения тех сообщений, которые имеют ценность для обнаружения инцидентов и подлежат нормализации в ходе обработки.

Шаг 3. В соответствии с рассмотренными выше методами разработать правила нормализации событий, определенных на шаге 2.

Шаг 4. В соответствии с рассмотренными выше методами определить необходимость и разработать способы агрегации сообщений, определенных на шаге 3. В случае отсутствия необходимости агрегации для конкретного типа сообщений данный шаг может быть пропущен.

Шаг 5. Основываясь на рассмотренных выше методах, разработать алгоритмы и правила корреляций нормализуемых событий. Поскольку именно в результате обнаружения корреляций создаются записи об инцидентах, то именно они и будут являться данными, подлежащими визуализации. При этом необходимо определить принципиальные схемы событий. Методика основывается на положении, что в каждом событии можно выделить следующие элементы:

– субъект: инициатор взаимодействия, о котором повествует событие;

– объект: основная сущность, описываемая событием;

– источник: наблюдатель события, фактически ИТ-актив, сформировавший и приславший это сообщение.

Исходя из этого можно выделить следующие схемы событий с точки зрения взаимодействия субъекта и объекта.

1) В событии не содержится информации о каком-либо взаимодействии – источник передает информацию о состоянии объекта (рис. 3а).

2) В событии есть информации только об объекте, но субъект подразумевается контекстом события (рис. 3б).

3) В событии объект выполняет действие над самим собой (рис. 3в).

4) В событии имеется информация о субъекте и объекте, а также их взаимодействии (рис. 3г).

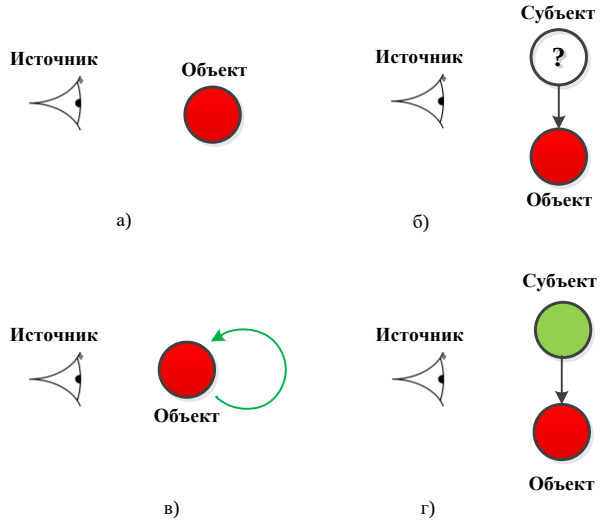

Рис. 3. Схемы событий с точки зрения взаимодействия субъекта и объекта

Шаг 6. Обеспечить накопление событий и записей об инцидентах в хранилище, обеспечивающем возможность работы с большими данными и предоставляющим функции поиска, в том числе поиска с нечеткими условиями.

Шаг 7. Определить возможные для использования графические модели [5] и скорректировать их с учетом сценария работы SIEM-системы. При определении моделей возможно использование критериев эффективности восприятия подсистемы визуализации [5].

Шаг 8. В соответствии с определенными на шаге 7 моделями реализовать визуализацию данных об инцидентах, извлекаемых из хранилища с помощью средств поиска, посредством использования программных компонентов или отдельных продуктов.

5. Апробация методики

Провести сравнение предложенного методического аппарата с исследованиями других ученых [3-5] на основе количественных характеристик не представляется возможным ввиду их различия в части основных положений методик. В связи с этим был проведен эксперимент по применению предложенной методики для решения задачи визуализации событий информационной безопасности в сети лаборатории безопасности интернета вещей ТУСУР.

Ранее было установлено, что в результате логирования событий, происходящих на IoT-устройствах, могут формироваться большие объемы данных. Для практической реализации предложенной авторами методики необходимо решить проблемы обработки больших данных. В [5,8] представлены обзоры рынка современных решений в области визуализации данных, среди которых стоит отметить продукт Elasticsearch. Это свободно распространяемый движок, предоставляющий распределенное аналитическое ядро поисковой системы. Используется в составе с Logstash [5] и Kibana [5]. Для взаимосвязи компонентов используется платформа обмена сообщениями RabbitMQ [8]. На рис. 4 представлена архитектура взаимодействия компонентов для организации возможности поиска, анализа и визуализации больших данных.

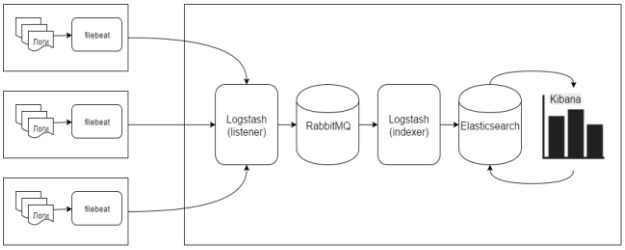

Рис. 4. Архитектура взаимосвязи компонентов

На базе стационарного компьютера был развернут стенд, имеющий следующие характеристики: 4-х ядерный процессор с тактовой частотой 3,1 ГГц; 16 Гб ОЗУ, на котором установлен стек ELK в составе: Elasticsearch 6.2.0, Logstash 6.2.0, Kibana 6.2.0. Стенд был развернут на базе операционной системы CentOS 7. Средняя скорость индексирования в Elasticsearch составила 15000 пакетов в секунду при средней загрузке каждого из ядер процессора 50%. Однако, поскольку основной целью эксперимента была оценка времени поиска среди индексированных сообщений, то были предприняты попытки отправки запросов в Elasticsearch для поиска среди сообщений слова из 6 букв. Среднее время поиска составило 1,02 мс.

Несмотря на достигнутые результаты в части поиска данных, были выявлены следующие проблемы при обработке больших данных, формирующихся в ходе логирования событий IoT-устройств. Во-первых, стек ELK критичен к ошибкам типа «OutOfMemory», что приводит к частым перебоям в ходе его использования. Кроме того,

Работа опубликована при финансовой поддержке РФФИ, грант 18-07-20045\18

высокая скорость поиска в больших объемах данных, достигаемая в первую очередь за счет документоориентированности сопровождается низким коэффициентом восстановления работоспособности в случае перебоев, что зачастую приводит к безвозвратной потере данных. С точки зрения безопасности ElasticSearch не имеет возможности для авторизации.

Одним из наиболее практичных подходов является применение правил, ограничивающих глубину корреляции и разделение базы событий на онлайн и архивную части. Например, события, произошедшие за последние сутки, хранятся в онлайн базе, по истечению таймера помещаются в архивную часть. Для работы с большим объемом данных применяются различные специализированные поисковые движки и инструменты визуализации. В следующих этапах исследования будут проведены эксперименты по использованию данного продукта в качестве поискового инструмента к различным реляционным СУБД для повышения стабильности работы и сохранности обрабатываемых данных. В некоторых публикациях [3-5] представлены данные смежных экспериментов. Ниже рассмотрены попытки сравнить полученные данные.

1. Аппаратные ресурсы. В большинстве рассмотренных примеров [3-5] используется распределение нагрузки на потоки в кластере из нескольких узлов (серверов). В данном случае использовались виртуальные машины на базе одного физического стенда, характеристики которого указаны выше.

2. Скорость индексирования и анализа используемых данных. В связи с невозможностью проведения экспериментов при равных условиях на идентичном оборудовании [3-5] проведение какой-либо количественной оценки по данному параметру не представляется возможным. Однако, такие сравнения вероятно будут проведены на следующих этапах исследования.

3. Механизм обработки данных. В эксперименте, как и в большинстве смежных исследований [3-5], применялась потоковая обработка.

4. Пул задач. Все рассмотренные задачи в смежных исследованиях имеют конкретную постановку и получить результаты их сравнения в численном виде не представляется возможным. Однако, на следующем этапе исследования будет возможно применение предлагаемого методического обеспечения для решения различных задач и получения количественных оценок будущих результатов.

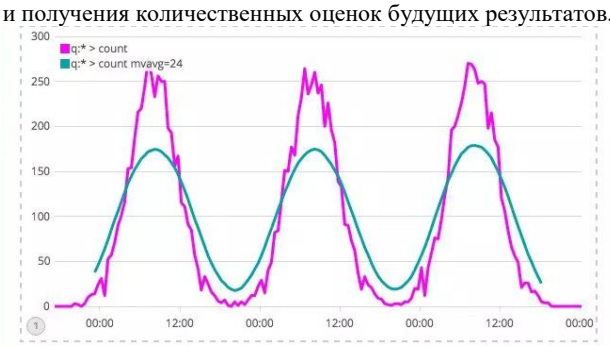

Рис. 5. Пример визуализации с помощью Kibana

На рисунке 5 представлен пример визуализации данных с помощью Kibana.

6. Заключение

Обобщение методических и практических наработок авторов в части организации сбора и обработки данных представлено в виде методики визуализации как средства совершенствования методов управления инцидентами. Информация о состоянии защищенности ИТ-активов в инфраструктуре интернета вещей может быть использована для непосредственного анализа, обнаружения инцидентов, а также их расследования и принятия решений.

Проведенные эксперименты показали, что применение методов корреляции является одним из инструментов совершенствования методов управления инцидентами в инфраструктуре Интернета вещей.

7. Благодарности

Работа выполнена при финансовой поддержке РФФИ (проект № 16-47-700350 р_а).

8. Литература

[1] *Abomhara M., Kien G.M.* Cyber security and the internet of things: vulnerabilities, threats, intruders and attacks // Journal of Cyber Security. – 2015. – Vol. 4. – Pp. 65–88.

[2] MaxPatrol SIEM [Электронный ресурс]. – Режим доступа: https://www.ptsecurity.com/ru-ru/products/mpsiem/?utm_source=slider (дата обращения: 23.06.2018).

[3] Dumitras T., Shou D.Toward a Standard Benchmark for Computer Security Research: the Worldwide Intelligence NetworkEnvironment (WINE) // Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS'11). 2011. pp. 89–96.

[4] Shu X., Smiy J., Yao D., Lin H. Massive Distributed and Parallel Log Analysis For Organizational Security // IEEE Globecom Workshops. December 2013. pp. 194–199

[5] Котенко И.В., Кулешов А.А., Ушаков И.А. Система сбора, хранения и обработки информации и событий безопасности на основе средств Elastic Stack/ [Электронный ресурс]. – Режим доступа http://proceedings.spiiras.nw.ru/ojs/index.php/sp/article/view/3590/2090 (дата обращения: 23.06.2018)/

[6] Милославская Н., Толстой А., Бирюков А. Визуализация информации при управлении информационной безопасностью информационной инфраструктуры организации //Научная визуализация. 2014. №2. [Электронный ресурс]. – Режим доступа http://sv-journal.org/2014-2/06/ru/index.php?lang=ru (дата обращения: 23.06.2018).

[7] Семёнов Д.П. Визуализация процессов информационной безопасности // Актуальные проблемы авиации и космонавтики. 2017. №13. [Электронный ресурс]. – Режим доступа: https://cyberleninka.ru/article/n/vizualizatsiya-protsessov-informatsionnoy-bezopasnosti (дата обращения: 23.06.2018).

[8] Корреляция SIEM. [Электронный ресурс]. – Режим доступа: https://www.securitylab.ru/analytics/431459.php (дата обращения: 23.06.2018)

Об авторах

Исхаков Сергей Юнусович, к.техн.н., доцент кафедры безопасности информационных систем Томского государственного университета систем управления и радиоэлектроники iskhakov.sy@gmail.com.

Исхаков Андрей Юнусович, к.техн.н., доцент кафедры безопасности информационных систем Томского государственного университета систем управления и радиоэлектроники iskhakovandrey@gmail.com.

Мещеряков Роман Валерьевич, д.техн.н., профессор кафедры безопасности информационных систем Томского государственного университета систем управления и радиоэлектроники mrv@ieee.org.

The publication of the conference proceedings was supported by the RFBR, grant No. 18-07-20045\18